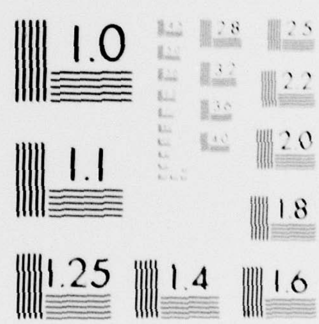
AD-A078 622 TEXAS UNIV AT AUSTIN CENTER FOR CYBERNETIC STUDIES F/G 12/1
A LINEAR PROGRAMMING APPROACH TO A SIMPLE LINEAR REGRESSION PRO--ETC(U)
SEP 79 R D ARMSTRONG , M T KUNG N00014-75-C-0616
UNCLASSIFIED CCS-351 NL

1 OF 1
AD
A078622

END
DATE
FILMED
1-80
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

CENTER FOR CYBERNETIC STUDIES

The University of Texas
Austin, Texas 78712

(12)

Research Report CCS 351

A LINEAR PROGRAMMING APPROACH TO A SIMPLE LINEAR REGRESSION PROBLEM WITH LEAST ABSOLUTE VALUE CRITERION

by

R. D. Armstrong
M. T. Kung

September 1979

This research was partly supported by Project NR047-021, ONR Contracts N00014-75-C-0616 and N00014-75-C-0569 with the Center for Cybernetic Studies, The University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the United States Government.

CENTER FOR CYBERNETIC STUDIES

A. Charnes, Director
Business-Economics Building, 203E
The University of Texas at Austin
Austin, TX 78712
(512) 471-1821

A LINEAR PROGRAMMING APPROACH TO A SIMPLE LINEAR REGRESSION PROBLEM WITH LEAST ABSOLUTE VALUE CRITERION

R. D. ARMSTRONG
Associate Professor
Operations Research and Statistics
Graduate School of Business
University of Texas
Austin, Texas 78712

and

M. T. KUNG
Research Associate
Center for Cybernetic Studies
University of Texas
Austin, Texas 78712

ABSTRACT

This paper presents a linear programming approach to solve simple linear regression problems with the least absolute value criterion. The solution technique uses linear programming with an extended minimum ratio rule. A computational study indicates the efficiency of the algorithm.

KEY WORDS

Simple Linear Regression Problem

Least Absolute Value Regression Problem

Goal Programming

Linear Programming

1. Introduction

The simple linear regression problem arises from a fundamental model of statistical analysis. The model consists of an independent (also known as predictor) random variable which is used to determine the value of the dependent (or response) random variable. The simple linear regression fit has been widely used in statistical and economic forecastings. The simple linear regression probelm is to find the linear equation which will fit the data comprising of these two variables.

The simple regression problem with a least absolute value criterion has the following form.

$$\text{(1)} \qquad \underset{(\alpha,\ \beta)}{\text{Minimize}} \ \sum_{i=1}^{n} | y_i - \alpha - x_i\beta|$$

where (x_i, y_i), i=1, 2, ..., n are the observed values.

2. Algorithm

Problem (1) is equivalent to the following linear programming problem [see 4]:

$$\text{Minimize} \ \sum_{i=1}^{n} (P_i + N_i)$$

$$\text{(2)} \qquad \text{subject to } \alpha + x_i\beta + P_i - N_i = y_i, \quad i=1, 2, \ldots, n$$

$$P_i \geq 0 \text{ and } N_i \geq 0, \quad i= 1, 2, \ldots, n$$

where P_i and N_i are, respectively, the positive and negative deviation associated with the i-th observation.

The dual problem of (2) is:

$$\text{Maximize} \ \sum_{i=1}^{n} \pi_i y_i$$

$$(3) \qquad \text{subject to} \quad \sum_{i=1}^{n} \pi_i = 0$$

$$\sum_{i=1}^{n} \pi_i x_i = 0$$

$$-1 \leq \pi_i \leq 1, \; i=1, 2, \ldots, n$$

We shall exploit the structure of (3) to solve this problem with a dual simplex algorithm. This process is identical to solving the problem (2) with the primal simplex algorithm. The following presentation develops a special purpose algorithm using the revised simplex method on the primal problem (2) with a multiple pivot strategy. This strategy enables the method to perform a pivot through several bases in one iteration.

Initially we choose two observations (x_c, y_c) and (x_d, y_d) such that $x_c \neq x_d$. Hence, the current basis for the LP problem is:

$$X_B = \begin{pmatrix} 1 & x_c \\ 1 & x_d \end{pmatrix}$$

The current right hand side is:

$$Y_B = \begin{pmatrix} y_c \\ y_d \end{pmatrix} .$$

By the adjoint form of the inverse, the initial basis inverse is given as follows:

$$X_B^{-1} = \frac{1}{x_c - x_d} \begin{pmatrix} -x_d & x_c \\ 1 & -1 \end{pmatrix}$$

The solution of (2) can be calculated:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = X_B^{-1} Y_B$$

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{x_c - x_d} \begin{pmatrix} -x_d & x_c \\ 1 & -1 \end{pmatrix} \begin{pmatrix} y_c \\ y_d \end{pmatrix}$$

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{x_c - x_d} \begin{pmatrix} -x_d y_c + x_c y_d \\ y_c - y_d \end{pmatrix}$$

Let NB represent the index set of the nonbasic rows. The deviations for (2), or the reduced costs for (3), are given by:

$$d_i = y_i - \alpha - \beta x_i$$

Define $\pi_i = \text{sign}(d_i)$, $i \in NB$. In the computer code, the value assigned to π_i when $i \in NB$ and $d_i = 0$ is arbitrarily defined to be +1 and, thereafter, the value is determined by the steps of the algorithm. The situation where $d_i = 0$ and $i \in NB$ corresponds to the case of degeneracy in linear programming and can be resolved as described by Charnes [3]. The details of this procedure will not be discussed here.

The nonbasic dual variables are either +1 or -1 depending on the sign of d_i, $i \in NB$. The values of the basic dual variables π_c and π_d are:

$$\begin{pmatrix} \pi_c \\ \pi_d \end{pmatrix} = \left(-\sum_{i \in NB} \pi_i, \quad -\sum_{i \in NB} \pi_i x_i \right) X_B^{-1}$$

$$\begin{pmatrix} \pi_c \\ \pi_d \end{pmatrix} = \left(-\sum_{i \in NB} \pi_i, \ -\sum_{i \in NB} \pi_i x_i \right) \frac{1}{x_c - x_d} \begin{pmatrix} -x_d & x_c \\ 1 & -1 \end{pmatrix}$$

$$\begin{pmatrix} \pi_c \\ \pi_d \end{pmatrix} = \begin{pmatrix} \dfrac{1}{x_c - x_d} \left(x_d \sum_{i \in NB} \pi_i - \sum_{i \in NB} \pi_i x_i \right) \\ \dfrac{1}{x_c - x_d} \left(-x_c \sum_{i \in NB} \pi_i + \sum_{i \in NB} \pi_i x_i \right) \end{pmatrix}$$

Since this is a primal algorithm, the optimality condition for (2) is dual feasibility, namely, $-1 \leq \pi_i \leq 1$, $i = c, d$.

If $|\pi_c| > 1$, the basic dual variable, π_c, will leave the basis. If $|\pi_c| \leq 1$, and $|\pi_d| > 1$, the algorithm interchanges the indices c and d.

Thus the variable to leave the basis will always be π_c. Define ρ = sign (π_c). The value of ρ indicates if π_c is to be increased or decreased. A value, $\rho = +1$ if π_c is to be decreased and $\rho = -1$ if π_c is to be increased.

The algorithm then determines a nonbasic dual variable to enter the basis. The procedure is to take the minimum value from a list of ratios:

(4)
$$\Theta_i = \frac{y_i - \alpha - \beta x_i}{\rho\xi_i} \qquad \text{for } \pi_i \rho \xi_i > 0, \qquad i \varepsilon NB,$$

where $\xi_i = \dfrac{x_i - x_d}{x_c - x_d}$, $i \varepsilon NB$.

Suppose θ_s is the smallest ratio value, then π_s, $s \varepsilon NB$, is the nonbasic dual variable to be examined. Firstly, the algorithm checks if π_s will enter the basis at a dual feasible level by the following criterion:

(5)
$$|\pi_c| - 2|\xi_s| \leq 1.$$

If condition (5) is satisfied, π_s will enter the basis and π_s will be dual feasible. The algorithm assigns c to be the current value of d. Then, d is given the value of s. For example, if π_3 and π_5 are currently basic, and π_3 is leaving the basis, and if the incoming nonbasic variable is π_{11}, the new basic variables will be π_5 and π_{11}, respectively.

On the other hand, if condition (5) is not satisfied, π_s will remain as a nonbasic dual variable because by bringing π_s into the basis, π_s will still be dual infeasible for (2). Rather, π_s will switch from its current bound value to its opposite bound value. Moreover the value of the basic dual variable π_c will be increased (or decreased) by

$2\rho|\xi_s|$. The algorithm then eliminates θ_s from the list of ratios

and examines the next possible candidate to enter the basis from (4).

The algorithm repeats the above procedure until $-1 \leq \pi_i \leq +1$, $i=c,d$ is satisfied. It may be noted that to check this condition after the first iteration only π_c need be examined.

3. Steps of the algorithm

In this section, we summarize the algorithm by giving a step-by-step description. New notation, such as D, T_1, T_2, and T, are introduced to make the algorithm easier to follow.

1. Initialization:

Choose two observations (x_c, y_c) and (x_d, y_d) such that $x_c \neq x_d$.

Set $D = \dfrac{1}{x_c - x_d}$

$$\alpha = (-x_d y_c + x_c y_d)D$$

$$\beta = (y_c - y_d)D$$

$$\pi_i = \text{sign}(y_i - \alpha - \beta_i) \quad i \neq c,d$$

$$\pi_c = \pi_d = 0$$

$$T_1 = \sum_{i=1}^{n} \pi_i$$

$$T_2 = \sum_{i=1}^{n} \pi_i x_i$$

$$T = (x_d T_1 - T_2)D$$

2. If $|T| > 1$, go to step 4. Otherwise, set $D = -D$, interchange π_c and π_d by setting $u \leftarrow c$, $c \leftarrow d$, $d \leftarrow u$, continue.

3. $T = (T_2 - x_d T_1)D$

If $|T| > 1$, go to step 4. Otherwise, stop. The current solution is optimal.

4. Set $\rho = \text{sign}(T)$

Determine the minimum from the following list of ratios:

$$\Theta_s = \frac{y_i - \alpha - \beta x_i}{\rho(x_i - x_d)D} \quad \text{for } \rho(x_i - x_d)D\pi_i > 0$$

Let Θ_s be the smallest ratio determined by the minimum ratio test.

5. If $|T| - 2|(x_s - x_d)D| \leq 1$ go to step 7. Otherwise, proceed to step 6.

6. $T = T - 2\rho|(x_s - x_d)D|$

$\pi_s \leftarrow -\pi_s$

$T_1 \leftarrow T_1 + 2\pi_s$

$T_2 \leftarrow T_2 + 2\pi_s x_s$

Eliminate Θ_s from the list of ratios and go to step 4.

7. $\pi_c = -\rho$

$T_1 \leftarrow T_1 - \rho - \pi_s$

$T_2 \leftarrow T_2 - \rho x_c - \pi_s x_s$

$\pi_s = 0$

π_d will replace π_c and π_s will replace π_d by setting $c \leftarrow d$, $d \leftarrow s$.

Set $D = \dfrac{1}{x_c - x_d}$

$\alpha = (-x_d y_c + x_c y_d)D$

$\beta = (y_c - y_d)D$

Go to step 3.

4. Computation Experience

A computational study was carried out to compare the FORTRAN code, LONESL, developed by Sadovski [5] and the FORTRAN code, SIMLP, [1] utilizing the algorithm presented there. Both are special purpose codes

designed expressly to provide least absolute value estimates for a simple linear regression model. The University of Texas CDC 6600 was used in this study. The observations have been drawn from various uniform and normal distributions using a random number generator. The results of the study are summarized in Table 1.

Our study has indicated that this specialization of the linear programming approach first developed by Barrodale and Roberts [2] is uniformly faster than the Sadovski's approach on all problem sizes. In problems with more than 300 observations, the SIMLP is approximately 50 times faster than LONESL. Considerably less storage is required in SIMLP when compared to the Barrodale and Roberts' code and approximately the same amount when compared to the Sodovski's code. Also, Sposito [6] has shown that LONESL may not always converge, while SIMLP utilizes the convergent properties of linear programming theory. Another feature of SIMLP is that there is no accumulative roundoff error present since all necessary values are recalculated from the original data at each iteration.

5. Conclusion

This paper presents a special purpose algorithm to solve simple least absolute value regression problems. The approach utilizes the characteristics and convergent properties of linear programming. With the addition of the multiple pivot strategy in linear programming, the simple least absolute value regression is solved efficiently. From the computational results, it is shown that the code presented here is superior to the published code of Sadovski, in terms of solution time. A listing of the computer code will be found in [1].

Table 1 (A Comparison between SIMLP and LONESL)

Number of observations	SIMLP		LONESL	
	Time (CPU milliseconds)	Number of iterations	Time (CPU milliseconds)	Number of iterations
50	7	3	57	2
100	32	6	407	4
150	46	5	953	4
200	102	5	1647	4
250	84	2	3230	5
300	70	3	8373	9
350	262	3	6367	5
400	265	3	16296	10
450	212	5	8341	4
500	184	3	15374	6

References

1. Armstrong, R.D. and M. T. Kung, (1978), "Least Absolute Value Estimates for a Simple Linear Regression Problem," Applied Statistics, 27, #3, 363-366.

2. Barrodale, I. and Roberts, F. D. K., (1973), "An Improved Algorithm for Discrete L_1 Linear Approximation," SIAM Journal on Numerical Analysis, 10, 839-848.

3. Charnes, A. (1952), "Optimality and Degeneracy in Linear Programming," Econometrica, 20, 160-170.

4. Charnes, A., Cooper, W. W. and Ferguson, R., (1955), "Optimal Estimation of Executive Compensation by Linear Programming," Management Science, 2, 138-151.

5. Sadovski, A. N., (1974), "Algorithm AS74. L_1-Norm Fit of a Straight Line," Applied Statistics, 23, 224-248.

6. Sposito, V. A., (1976), "A Remark on Algorithm AS74: L_1-Norm Fit of a Straight Line," Applied Statistics, 25, 96-97.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) The Center for Cybernetic Studies The University of Texas at Austin	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

6 A Linear Programming Approach to a Simple Linear Regression Problem with Least Absolute Value Criterion

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

10 R.D. Armstrong [redacted] M.T. Kung

6. REPORT DATE 11 Sep [redacted] 79	7a. TOTAL NO. OF PAGES 11	7b. NO. OF REFS 6

8a. CONTRACT OR GRANT NO. 15 N00014-75-C-0616, [redacted] N00014-75-C-0569 b. PROJECT NO. NR047-071 and NR047-021 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Center for Cybernetic Studies 9 Research Rep[redacted]t, 351 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) 12 15

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research (Code 434) Washington, DC

13. ABSTRACT

This paper presents a linear programming approach to solve simple linear regression problems with the least absolute value criterion. The solution technique uses linear programming with an extended minimum ratio rule. A computational study indicates the efficiency of the algorithm.

14 225-351

406 197

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

A-31408

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Simple Linear Regression Problem						
Least Absolute Value Regression Problem						
Goal Programming						
Linear Programming						

DD FORM 1 NOV 65 1473 (BACK)
S/N 0102-014-6800

Unclassified
Security Classification

A-31409